Responsible and Safe AI Systems

RSA Week 1

1) According to the risk decomposition framework, which combination of factors would result in the HIGHEST risk from an AI system deployed in a critical infrastructure setting? ***1 point***

- Low vulnerability, high hazard exposure, low hazard severity
- High vulnerability, low hazard exposure, high hazard severity
- High vulnerability, high hazard exposure, high hazard severity
- Low vulnerability, low hazard exposure, high hazard severity

Yes, the answer is correct.
Score: 1
Accepted Answers:
High vulnerability, high hazard exposure, high hazard severity

2) The concept of treacherous turns in AI systems refers to: ***1 point***

- AI systems making computational errors during complex calculations
- AI systems behaving differently once they reach sufficient intelligence
- AI systems being hacked by malicious actors
- AI systems consuming too much computational power

Yes, the answer is correct.
Score: 1
Accepted Answers:
AI systems behaving differently once they reach sufficient intelligence

3) In the context of AI race dynamics, what is the primary concern regarding competitive pressure between nations and corporations? ***1 point***

- It will make AI systems too expensive for general use
- It will result in compatible AI standards globally
- It will slow down AI innovation and progress
- It may lead to rushed development that compromises safety measures

Yes, the answer is correct.
Score: 1
Accepted Answers:
It may lead to rushed development that compromises safety measures

4) The "Swiss cheese model" mentioned in organizational risks suggests that: ***1 point***

- Organizations should have a single, very strong safety measure
- Safety measures should be implemented randomly across the organization
- Multiple layers of defense compensate for individual weaknesses
- Safety measures are unnecessary if the AI system is well-designed

Yes, the answer is correct.
Score: 1
Accepted Answers:
Multiple layers of defense compensate for individual weaknesses

5) Which scenario best illustrates the concept of proxy gaming? ***1 point***

- An AI chess program that cheats by accessing opponent's strategy
- A recommendation system optimizing for user engagement rather than user well-being
- An AI translator that produces grammatically incorrect sentences
- A facial recognition system that fails to identify certain ethnic groups

Yes, the answer is correct.
Score: 1
Accepted Answers:
A recommendation system optimizing for user engagement rather than user well-being

6) A factory robot confuses a human worker for a box of vegetables and pushes the person, resulting in death." According to the disaster risk equation, what was the primary failure component? ***1 point***

- Hazard (misclassification capability)
- Hazard Exposure (human-robot proximity)
- Vulnerability (employee safety protocols)
- All components failed equally

Yes, the answer is correct.
Score: 1
Accepted Answers:
Hazard (misclassification capability)

7) According to the risk taxonomy presented, malicious use of AI differs from rogue AI primarily in that: **1 point**

- (●) Malicious use involves intentional harmful deployment by humans, while rogue AI acts independently
- ( ) Malicious use only affects cybersecurity, while rogue AI affects all domains
- ( ) Malicious use is easier to detect than rogue AI behavior
- ( ) Malicious use requires more advanced AI capabilities than rogue AI

Yes, the answer is correct.
Score: 1
Accepted Answers:
Malicious use involves intentional harmful deployment by humans, while rogue AI acts independently

8) Deceptive Alignment in AI systems is: **1 point**

- ( ) AI systems that are openly hostile to humans
- (●) AI systems that appear to be following instructions but are actually pursuing different goals
- ( ) AI systems that cannot understand human language properly
- ( ) AI systems that work too slowly to be effective

Yes, the answer is correct.
Score: 1
Accepted Answers:
AI systems that appear to be following instructions but are actually pursuing different goals

9) How do you identify and avoid hazards in ML systems according to the disaster risk equation framework? **1 point**

- ( ) Alignment
- (●) Robustness
- ( ) Monitoring
- ( ) Systemic Safety

No, the answer is incorrect.
Score: 0
Accepted Answers:
Monitoring

10) Red teaming in AI safety primarily serves to: **1 point**

- ( ) Accelerate model training
- (●) Identify system vulnerabilities
- ( ) Improve computational efficiency
- ( ) Reduce inference latency

Yes, the answer is correct.
Score: 1
Accepted Answers:
Identify system vulnerabilities

11) Which technique is most effective for detecting deceptive alignment? **1 point**

- ( ) Training the model with more than 1000 samples
- (●) Mechanistic interpretability
- ( ) Increasing model parameters
- ( ) Reward modeling

Yes, the answer is correct.
Score: 1
Accepted Answers:
Mechanistic interpretability

12) RoBERTa succeeds in reasoning tasks where BERT fails due to: **1 point**

- ( ) Better tokenization
- (●) Emergent capabilities from scaling
- ( ) Improved attention mechanisms
- ( ) Larger vocabulary size

Yes, the answer is correct.
Score: 1
Accepted Answers:
Emergent capabilities from scaling

RSA Week 2

Assignment submitted on 2025-08-04, 23:24 IST

1) Which of the following is true? *1 point*

- Extremistan has thin tails while Mediocristan has long tails
- Mediocristan distributions are harder to predict than Extremistan
- (selected) In Extremistan, the total is determined by a few large events with tyranny of the accidental
- Extremistan has mild randomness while Mediocristan has wild randomness

Yes, the answer is correct.
Score: 1
Accepted Answers:
In Extremistan, the total is determined by a few large events with tyranny of the accidental

2) What is the key difference between covariate shift and concept shift in distribution shifts? *1 point*

- Covariate shift changes P(y|x) while concept shift changes P(x)
- (selected) Covariate shift changes P(x) while P(y|x) remains constant, concept shift changes P(y|x) while P(x) remains constant
- Both change P(x) and P(y|x) simultaneously
- Covariate shift affects labels while concept shift affects features

Yes, the answer is correct.
Score: 1
Accepted Answers:
Covariate shift changes P(x) while P(y|x) remains constant, concept shift changes P(y|x) while P(x) remains constant

3) In the AugMix methodology, what is the primary advantage over uncontrolled random augmentations? *1 point*

- (selected) It uses skip connections to keep images recognizable while applying diverse augmentations
- It requires less computational power
- It only applies single augmentations instead of multiple
- It focuses on geometric transformations only

Yes, the answer is correct.
Score: 1
Accepted Answers:
It uses skip connections to keep images recognizable while applying diverse augmentations

4) In the context of adversarial attacks, what does "transferability" specifically refer to? *1 point*

- The ability to transfer attacks from one domain to another
- The ability to transfer defenses across different architectures
- The ability to convert white-box attacks to black-box attacks
- (selected) The ability of adversarial examples crafted for one model to work on other models

Yes, the answer is correct.
Score: 1
Accepted Answers:
The ability of adversarial examples crafted for one model to work on other models

5) Black Swan lies in which of the following categories? *1 point*

- Known Knowns
- Known Unknowns
- Unknown Knowns
- (selected) Unknown Unknowns

Yes, the answer is correct.
Score: 1
Accepted Answers:
Unknown Unknowns

6) Which of the following are valid approaches for defending against adversarial attacks? (Select all that apply) *1 point*

- [x] Data augmentation techniques
- [x] Adversarial training using adversarial examples during training
- [x] Using more data and larger models
- [ ] Reducing model complexity to avoid overfitting
- [ ] Adversarial pretraining on larger datasets like ImageNet

Partially Correct.
Score: 0.75
Accepted Answers:
Data augmentation techniques
Adversarial training using adversarial examples during training
Using more data and larger models
Adversarial pretraining on larger datasets like ImageNet

7) In the RLHF optimization objective, why is a KL-divergence penalty term added to the reward maximization? **1 point**

- To prevent the model from generating repetitive outputs
- To ensure the model stays close to the original pretrained model
- To improve the computational efficiency of the training process
- To increase the diversity of generated samples

Yes, the answer is correct.
Score: 1
Accepted Answers:
To ensure the model stays close to the original pretrained model

8) What does "reward hacking" specifically refer to in the context of RLHF? **1 point**

- Humans providing incorrect feedback to manipulate the system
- External attackers compromising the reward model
- The reward model overfitting to the training data
- The model finding ways to maximize the reward function without achieving the intended behavior

Yes, the answer is correct.
Score: 1
Accepted Answers:
The model finding ways to maximize the reward function without achieving the intended behavior

9) Identify the equations that can lead to a long-tailed distribution. **1 point**

- Idea * student * resources * time
- Idea * student + resources * time
- Idea + student + resource + time
- Idea - student * resource - time

Yes, the answer is correct.
Score: 1
Accepted Answers:
*Idea * student * resources * time*

10) What is the primary advantage of using pairwise comparisons over direct scalar ratings in human feedback collection? **1 point**

- Pairwise comparisons are faster to collect
- They require fewer human annotators
- Human judgments are noisy and miscalibrated, but pairwise comparisons are more reliable
- They provide more granular feedback information

Yes, the answer is correct.
Score: 1
Accepted Answers:
Human judgments are noisy and miscalibrated, but pairwise comparisons are more reliable

11) What is a major challenge with using a single reward function in RLHF? **1 point**

- It is computationally expensive to optimize
- It cannot represent a diverse society of humans
- It requires too much training data
- It is unstable during training

Yes, the answer is correct.
Score: 1
Accepted Answers:
It cannot represent a diverse society of humans

12) What is Direct Preference Optimization (DPO) equivalent to in the context of RLHF component removal? **1 point**

- RLHF - Human Feedback
- RLHF - Reward Model
- RLHF - RL
- RLHF - Policy Optimization

No, the answer is incorrect.
Score: 0
Accepted Answers:
RLHF - Reward Model

RSA Week 3

Assignment submitted on 2025-08-07, 19:42 IST

1) What is the definition of “Machine Unlearning”? **1 point**

- (•) Removing the influences of training data from a trained model
- ( ) The ability of a machine learning model to adapt to new data
- ( ) A technique used to compress machine learning models
- ( ) The ability of a machine learning model to learn a variety of data

Yes, the answer is correct.
Score: 1
Accepted Answers:
Removing the influences of training data from a trained model

2) What might be some types of information that one might want to remove from model data? (Select all that apply.) **1 point**

- [x] Private data
- [x] Toxic or unsafe content
- [ ] Accurate information
- [ ] Model hyperparameter settings
- [x] Stale knowledge

Yes, the answer is correct.
Score: 1
Accepted Answers:
Private data
Toxic or unsafe content
Stale knowledge

3) What is GDPR’s Article 17 about in the context of Machine Learning? **1 point**

- ( ) Right to be remembered
- ( ) Right to be modified
- ( ) Right to be distributed
- (•) Right to be forgotten

Yes, the answer is correct.
Score: 1
Accepted Answers:
Right to be forgotten

4) What are the steps in the SISA approach? **1 point**

- ( ) Sampled, Isolated, Stopped, Aggregated
- ( ) Sharded, Imitate, Sliced, Annotated
- (•) Sharded, Isolated, Sliced, Aggregated
- ( ) Sampled, Imitate, Stopped, Annotated

Yes, the answer is correct.
Score: 1
Accepted Answers:
Sharded, Isolated, Sliced, Aggregated

5) What is Membership Inference Attack (MIA) used for? **1 point**

- ( ) To train LLMs faster
- ( ) To improve the models robustness
- ( ) To remove accurate data
- (•) To classify between training and unseen data

Yes, the answer is correct.
Score: 1
Accepted Answers:
To classify between training and unseen data

6) What is the role of differential privacy? **1 point**

- ( ) To improve model accuracy
- ( ) To make the model forget everything
- (•) To make models indistinguishable with/without certain data points
- ( ) To speed up training time

Yes, the answer is correct.
Score: 1
Accepted Answers:
To make models indistinguishable with/without certain data points

7) Which benchmarks are used to evaluate unlearning in LLMs? (Select all that apply.) **1 point**

- [x] TOFU
- [ ] GLUE
- [x] WMDP
- [ ] GUIDE

Yes, the answer is correct.
Score: 1
Accepted Answers:
TOFU
WMDP

8) Case: A hospital wants to share patient health data with research institutions to support public health studies. However, it must ensure that a patient's identity cannot be inferred. To achieve this, the hospital generates and sends aggregate statistics to the institute. Which form of unlearning is this? **1 point**

- ( ) Exact unlearning
- (•) Unlearning via differential privacy
- ( ) Just ask for unlearning
- ( ) Empirical Unlearning

Yes, the answer is correct.
Score: 1
Accepted Answers:
Unlearning via differential privacy

9) Which isn't a type of graph unlearning? **1 point**

- ( ) Node unlearning
- ( ) Edge unlearning
- (•) Label unlearning
- ( ) Node feature unlearning

Yes, the answer is correct.
Score: 1
Accepted Answers:
Label unlearning

10) Which methods are used for Node unlearning? (Select all that apply.) **1 point**

- [x] GraphEraser
- [x] GUIDE
- [ ] Projector
- [x] GraphdEditor
- [x] GNNDelete
- [x] MEGU

Yes, the answer is correct.
Score: 1
Accepted Answers:
GraphEraser
GUIDE
GraphdEditor
GNNDelete
MEGU

11) What are "hidden representations"? **1 point**

- ( ) The final model outputs
- ( ) Raw training data stored in memory
- (•) Intermediate activation vectors captured during model execution
- ( ) The model's loss values during training

Yes, the answer is correct.
Score: 1
Accepted Answers:
Intermediate activation vectors captured during model execution

RSA Week 4

Assignment submitted on 2025-08-19, 12:18 IST

1) As per the lecture, in the context of Machine Learning, what is the definition of 'bias'? **1 point**

- Systematic deviation from rationality and judgement.
- Systematic error in the collection, analysis, or interpretation of data.
- (selected) Systematic favoritism or discrimination towards certain groups/outcomes.
- Systematic behaviour when solving complex tasks.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Systematic favoritism or discrimination towards certain groups/outcomes.

2) Why did Microsoft's Tay chatbot become offensive shortly after it was launched? **1 point**

- (selected) It learned toxic behaviour from user interactions on Twitter.
- It was hacked by a rival company.
- It was trained on outdated information.
- It was programmed to be controversial for publicity.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It learned toxic behaviour from user interactions on Twitter.

3) As per the lecture, which of the following is a/are a category of bias? (Select all that apply.) **1 point**

- (selected) Gender
- (selected) Race
- Scientific Facts
- (selected) Profession
- Currency exchange rates

Yes, the answer is correct.
Score: 1
Accepted Answers:
Gender
Race
Profession

4) According to the ML Pipeline, what may be a source of bias? (Select all that apply.) **1 point**

- (selected) Annotators beliefs
- Hardware used for computation
- Balanced dataset
- (selected) Biased training data

Yes, the answer is correct.
Score: 1
Accepted Answers:
Annotators beliefs
Biased training data

5) What does the Legal Safety Score (LSS_β) represent? **1 point**

- The model's ability to predict legal outcomes based solely on accuracy.
- (selected) A metric combining fairness and accuracy using a β-weighted harmonic mean.
- A score based on the usage of legal jargon for marginalized groups.
- An evaluation metric used to define how well a model understands legal jargon.

Yes, the answer is correct.
Score: 1
Accepted Answers:
A metric combining fairness and accuracy using a β-weighted harmonic mean.

6) What do LLMs use to prevent harmful outputs? **1 point**

- Data augmentation
- (selected) Guardrails
- Faster GPUs
- Dropout layers

Yes, the answer is correct.
Score: 1
Accepted Answers:
Guardrails

7) Which of the following is true about bias? **1 point**

- It never exists.
- It sometimes exists.
- It only exists in American-created models.
- It always exists.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It always exists.

8) A researcher is evaluating a facial recognition model they helped develop. During testing, they select images where the model performs better, such as images with ideal lighting or frontal faces, while ignoring diverse or difficult cases (like low-light, non-white faces, or side angles). Which type of bias is this? **1 point**

- Reporting bias
- Sampling bias
- Experimenter's bias
- Historical bias

Yes, the answer is correct.
Score: 1
Accepted Answers:
Experimenter's bias

9) What is the issue with evaluating models only based on accuracy? **1 point**

- Accuracy only reflects hardware performance.
- Accuracy doesn't reveal bias.
- Accuracy checks for fairness.
- Accuracy changes the labelling.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Accuracy doesn't reveal bias.

10) Why might using Western-aligned datasets be problematic for the Indian demographic? **1 point**

- The datasets are too large which may slow down training time.
- They are written in a different language.
- They don't reflect Indian social/societal norms.
- They contain too many low-resolution images.

Yes, the answer is correct.
Score: 1
Accepted Answers:
They don't reflect Indian social/societal norms.

11) What is a 'stereotype'? **1 point**

- A factual statement that applies to all humans.
- A scientifically proven characteristic.
- A legal rule used to govern a society.
- A widely held belief about some group/entity.

Yes, the answer is correct.
Score: 1
Accepted Answers:
A widely held belief about some group/entity.

12) What does the CrowS-Pairs dataset contain? **1 point**

- Pairs of sentences that differ only in minimally distant social bias.
- Dialogues between humans and a chatbot.
- Pairs of biased and unbiased images.
- Code snippets with and without bugs.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Pairs of sentences that differ only in minimally distant social bias.

13) What is sampling bias? **1 point**

- When historical data reflects inequalities that existed in the world at that time.
- When data is collected from a completely random and diverse group.
- If proper randomization is not used during data collection.
- When a model builder keeps training a model until it produces a result that aligns with their original hypothesis.

Yes, the answer is correct.
Score: 1
Accepted Answers:
If proper randomization is not used during data collection.

RSA Week 5

Assignment submitted on 2025-08-25, 15:09 IST

1) What is considered an ideal Stereotype Score (ss)? **1 point**

- 0%
- 25%
- 50%
- 75%

Yes, the answer is correct.
Score: 1
Accepted Answers:
50%

2) What does SEAT stand for? **1 point**

- Standard Evaluation Assessment Test
- Semantic Evaluation Annotation Test
- Structured Embedding Accuracy Test
- Sentence Embedding Association Test

Yes, the answer is correct.
Score: 1
Accepted Answers:
Sentence Embedding Association Test

3) In counterfactual data augmentation (CDA), what is altered to rebalance the corpus? **1 point**

- Sentence Length
- Syntex
- Bias attribute words
- Vocabulary complexity

Yes, the answer is correct.
Score: 1
Accepted Answers:
Bias attribute words

4) Which characteristic makes a language model more likely to generate gender-neutral responses in text-to-text tasks? (Select all that apply.) **1 point**

- [x] Training on diverse and balanced datasets.
- [x] Use of bias-specific adapter modules.
- [ ] Conditioning outputs on explicit gender tokens.
- [ ] Reliance on pretrained token embeddings without fine-tuning.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Training on diverse and balanced datasets.
Use of bias-specific adapter modules.

5) Which toolkit is used to add programmable guardrails to LLM-based conversational applications like ChatGPT? **1 point**

- GPT-4
- NVIDIA NeMo
- CoDi
- MAFIA

Yes, the answer is correct.
Score: 1
Accepted Answers:
NVIDIA NeMo

6) Which of the following is the correct formula for Pointwise Mutual Information (PMI)? **1 point**

- $PMI(w_i, w_j) = \log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i)^2 \cdot c(w_j)^2}$
- $PMI(w_i, w_j) = \log_2 \frac{c(w_i) \cdot c(w_j)}{N \cdot c(w_i, w_j)}$
- $PMI(w_i, w_j) = \log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i) \cdot c(w_j)}$
- $PMI(w_i, w_j) = \log_2 \frac{c(w_i, w_j)}{N \cdot c(w_i) \cdot c(w_j)}$

Yes, the answer is correct.
Score: 1
Accepted Answers:
$PMI(w_i, w_j) = \log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i) \cdot c(w_j)}$

7) 'Useful fairness' couples which of the following? (Select all that apply.) *1 point*

- [x] Context awareness
- [ ] Bias Score
- [x] STS task performance
- [ ] Dataset diversity

No, the answer is incorrect.
Score: 0
Accepted Answers:
Bias Score
STS task performance

8) What is the key architectural idea behind the MAFIA model for effective debiasing? *1 point*

- (•) Fusing bias-specific adapters while keeping the base model the same.
- ( ) Replacing all model weights with debiased adapters.
- ( ) Dynamically routing inputs based on detected bias type.
- ( ) Using GANs to hallucinate fair outputs.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Fusing bias-specific adapters while keeping the base model the same.

9) Which of the following is true about proprietary models? *1 point*

- ( ) They are more neutral compared to CoDi and other open source models.
- (•) They are less neutral compared to CoDi and other open source models.
- ( ) They have more bias than CoDi and other open source models.
- ( ) They have the same neutrality as CoDi and other open source models.

No, the answer is incorrect.
Score: 0
Accepted Answers:
They are more neutral compared to CoDi and other open source models.

10) Which of the following are benchmark datasets commonly used to measure bias in language models? (Select all that apply.) *0 points*

- [x] Stereoset
- [x] CrowS-Pairs
- [ ] ImageNet
- [x] Bias-STS-S
- [ ] SQuAD

Yes, the answer is correct.
Score: 0
Accepted Answers:
Stereoset
CrowS-Pairs
Bias-STS-S

11) What is 'gender-bleaching' in the context of VLMs? *1 point*

- ( ) Improving the quality of input images.
- ( ) Turning all people in the input images white.
- ( ) Enhancing gender-specific features in input images.
- (•) Removing/Obscuring visual cues related to gender in input images.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Removing/Obscuring visual cues related to gender in input images.

12) Why might a single adapter for all bias types (like iDEBall) fail? *1 point*

- (•) Cannot effectively debias across all categories.
- ( ) Trains slower than other models.
- ( ) Requires more input data.
- ( ) Does not understand contextual information.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Cannot effectively debias across all categories.

RSA Week 6

Assignment submitted on 2025-09-03, 22:22 IST

1) What are the primary disadvantages of cryptographic solutions for privacy protection? Select all that apply. **1 point**

- [ ] Inference possibility remains uncertain in all scenarios
- [x] Utility reduces
- [x] It is expensive
- [ ] Security guarantees are absolute in all cases

No, the answer is incorrect.
Score: 0
Accepted Answers:
Inference possibility remains uncertain in all scenarios
It is expensive

2) Why do anonymization techniques often fail to provide adequate privacy protection? **1 point**

- ( ) They require excessive computational resources
- (•) Adversaries can leverage auxiliary databases for de-anonymization attacks
- ( ) They only work with structured data formats
- ( ) The anonymization process introduces too much noise

Yes, the answer is correct.
Score: 1
Accepted Answers:
Adversaries can leverage auxiliary databases for de-anonymization attacks

3) In the randomized response model, what does the parameter epsilon in e^{ε} represent? **1 point**

- (•) The percentage of truthfulness of the response that is allowed to be revealed
- ( ) The ratio between falsehood and randomness in responses
- ( ) The computational complexity of the algorithm
- ( ) The number of participants in the study

Yes, the answer is correct.
Score: 1
Accepted Answers:
The percentage of truthfulness of the response that is allowed to be revealed

4) What characterizes an ideal privacy model? **1 point**

- ( ) Maximum utility with zero privacy guarantees
- ( ) Perfect privacy with minimal utility
- (•) Balanced trade-off between utility and privacy requirements
- ( ) Complete data suppression for absolute privacy

Yes, the answer is correct.
Score: 1
Accepted Answers:
Balanced trade-off between utility and privacy requirements

5) Given: Xi represents truth, Yi represents randomized value of Xi, and $Zi = (Yi - (1/(1+e^{\varepsilon}))) \times (e^{\varepsilon}+1)/(e^{\varepsilon}-1)$, What is the expected value Zi? **1 point**

- ( ) $E[Zi] = Yi$
- (•) $E[Zi] = Xi$
- ( ) $E[Zi] = e^{\varepsilon} \times Xi$
- ( ) $E[Zi] = 0$

Yes, the answer is correct.
Score: 1
Accepted Answers:
$E[Zi] = Xi$

6) In the randomized response model, which statements are correct? Select all that apply. **0 points**

- [x] Privacy guarantee can be controlled by parameter ε
- [ ] Privacy and utility are independent of sample size
- [x] Higher ε values provide stronger utility
- [x] Utility guarantee scales with $\sqrt{n}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
Privacy guarantee can be controlled by parameter ε
Utility guarantee scales with $\sqrt{n}$

7) Consider datasets:X= {x1, x2, ..., xN} (truth), Y = {y1, y2, ..., yN} (revealed values). To derive better estimators Z = {z1, z2, ..., zN} from Y, What process is required? **1 point**

- (•) Removing bias from Y introduced through randomization
- ( ) Adding bias to Y removed through randomization
- ( ) Removing variance from Y introduced through randomization
- ( ) Adding variance to Y removed through randomization

Yes, the answer is correct.
Score: 1
Accepted Answers:
Removing bias from Y introduced through randomization

8) Which factors directly influence the privacy guarantee in differential privacy mechanisms? Select all that apply. **1 point**

- [x] Sensitivity of the query function
- [x] Magnitude of added noise
- [ ] Size of the dataset
- [x] Privacy parameter Epsilon

Yes, the answer is correct.
Score: 1
Accepted Answers:
Sensitivity of the query function
Magnitude of added noise
Privacy parameter Epsilon

9) Which trust scenario correctly describes the differential privacy model? **1 point**

- ( ) Trust the curator; Trust the world
- ( ) Do not trust the curator; Trust the world
- (•) Trust the curator; Do not trust the world
- ( ) Do not trust the curator; Do not trust the world

Yes, the answer is correct.
Score: 1
Accepted Answers:
Trust the curator; Do not trust the world

10) For fixed privacy levels, how do sample size requirements differ between Laplacian mechanism and Randomized response? **1 point**

- (•) Constant factor difference
- ( ) Exponential factor difference
- ( ) Logarithmic factor difference
- ( ) Quadratic factor difference

No, the answer is incorrect.
Score: 0
Accepted Answers:
Quadratic factor difference

11) Higher privacy guarantees can be achieved in which scenarios? Select all correct options. **1 point**

- [ ] When Epsilon parameter is increased
- [x] When noise magnitude is increased
- [ ] When inverse sensitivity is high
- [x] When variance in the mechanism is high
- [ ] When utility requirements are maximized

Yes, the answer is correct.
Score: 1
Accepted Answers:
When noise magnitude is increased
When variance in the mechanism is high

12) In the context of randomized response, what happens to utility as privacy guarantees increase? **1 point**

- ( ) Utility remains constant
- ( ) Utility increases proportionally
- (•) Utility decreases due to increased noise
- ( ) Utility becomes undefined

Yes, the answer is correct.
Score: 1
Accepted Answers:
Utility decreases due to increased noise

RSA Week 7

Assignment submitted on 2025-09-10, 11:30 IST

1) In approximate differential privacy, what role does the delta (δ) parameter play?(Notation same as in the lecture) **1 point**

- It ensures that epsilon differential privacy holds for all possible sets
- It allows for some rare events (set S) where epsilon differential privacy may not hold
- It reduces the amount of noise required in all cases
- It eliminates the need for the epsilon parameter

Yes, the answer is correct.
Score: 1
Accepted Answers:
It allows for some rare events (set S) where epsilon differential privacy may not hold

2) What is a key advantage of approximate differential privacy over standard differential privacy? **1 point**

- It provides stronger privacy guarantees
- It eliminates the need for noise addition
- It can increase utility
- It works only with categorical data

Yes, the answer is correct.
Score: 1
Accepted Answers:
It can increase utility

3) In approximate differential privacy, the Gaussian noise follows which pattern? **1 point**

- $N(0, \sigma)$ where $\sigma = \sqrt{(d \log(1/\delta))/(n+\varepsilon)}$
- $N(0, \sigma)$ where $\sigma = \sqrt{(d \log(1/\delta))/(\varepsilon)}$
- $N(0, \sigma)$ where $\sigma = d \log(1/\delta)/(n-\varepsilon)$
- $N(0, \sigma)$ where $\sigma = \sqrt{(d \log(1/\delta))/(n \cdot \varepsilon)}$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$N(0, \sigma)$ where $\sigma = \sqrt{(d \log(1/\delta))/(n \cdot \varepsilon)}$

N multiply e

4) What does Statistical Parity require for a fair classifier? **1 point**

- The classifier should have equal accuracy across all groups
- The probability of positive predictions should be equal across protected and non-protected groups
- The true positive rates should be identical for all demographic groups
- Individual similar cases should receive similar predictions

Yes, the answer is correct.
Score: 1
Accepted Answers:
The probability of positive predictions should be equal across protected and non-protected groups

5) For an ideal fair algorithm under Equality of Opportunity, what condition must be satisfied? **1 point**

- The overall prediction rates must be equal across groups
- The false positive rates of unprivileged and privileged groups should be equal
- The true positive rates of unprivileged and privileged groups should be equal
- The precision should be identical for both protected and non-protected groups

Yes, the answer is correct.
Score: 1
Accepted Answers:
The true positive rates of unprivileged and privileged groups should be equal

6) What is the Post-Processing property in differential privacy? **1 point**

- Any preprocessing of data before applying DP mechanisms maintains the privacy guarantee
- Privacy guarantees are strengthened when multiple post-processing steps are applied
- Any data-independent transformation applied to the output of a differentially private mechanism does not degrade its privacy guarantee
- Post-processing can be used to improve the privacy guarantee of any mechanism

Yes, the answer is correct.
Score: 1
Accepted Answers:
Any data-independent transformation applied to the output of a differentially private mechanism does not degrade its privacy guarantee

7) In a PCA analysis, if the reconstruction error for female data points is lower than for male data points, what does this indicate? ***1 point***

- The dataset is biased against females
- The dataset is biased against males
- Reconstruction error differences are due to random noise
- Male data is more correlated

Yes, the answer is correct.
Score: 1

Accepted Answers:
The dataset is biased against males

8) In the exponential mechanism to calculate the price to maximize the revenue, identify the correct statement in the scenario where 2 unequal prices result in the same revenue: ***1 point***

- Both prices have an unequal probability of being selected
- Both prices have an equal probability of being selected
- A higher price has a higher probability of being chosen due to normalisation
- A lower price has a higher probability of being chosen due to normalisation

Yes, the answer is correct.
Score: 1

Accepted Answers:
Both prices have an equal probability of being selected

9) In an ideal situation where the models are completely fair, the different parity values are: ***1 point***

- Approach 0
- 1
- Approach 1
- 0

No, the answer is incorrect.
Score: 0

Accepted Answers:
0

10) In a classifier, if a data point lies exactly on the decision boundary (hyperplane), what is the probability of it belonging to the positive class? ***1 point***

- Greater than 50%
- Less than 50%
- Equal to 50%
- Cannot be determined from the given information

Yes, the answer is correct.
Score: 1

Accepted Answers:
Equal to 50%

11) In Fair Logistic Regression, the equation P(M(x)=1|C=1) - P(M(x)=1|C=0) represents which fairness metric? ***1 point***

- Equality of Opportunity
- Statistical Parity
- Predictive Parity
- Individual Fairness

Yes, the answer is correct.
Score: 1

Accepted Answers:
Statistical Parity

RSA Week 8

Assignment submitted on 2025-09-17, 19:38 IST

1) In pixel-attribution methods, which statement best distinguishes perturbation-based approaches from gradient-based saliency? ***1 point***

- Perturbation-based methods compute input gradients; gradient-based methods fit local surrogate models.
- Perturbation-based methods are model-agnostic and computationally expensive; gradient-based methods use model internals and are faster.
- Perturbation-based methods always produce sparser explanations; gradient-based methods always produce denser explanations.
- Perturbation-based methods require access to intermediate activations; gradient-based methods are strictly black-box.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Perturbation-based methods are model-agnostic and computationally expensive; gradient-based methods use model internals and are faster.

2) Which behavior of a saliency map under layer-randomization is a red flag that the map is not capturing learned model structure? ***1 point***

- The saliency map remains largely unchanged after randomizing many layers.
- The saliency map changes dramatically as earlier layers are randomized.
- The saliency map's sign (positive/negative) flips while spatial structure remains.
- Absolute saliency values shrink while layout shifts.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The saliency map remains largely unchanged after randomizing many layers.

3) Which of the following is fully equivariant to translation and rotation: ***1 point***

- StyleGAN2
- ProtoPNet
- StyleGAN3
- None of the above

Yes, the answer is correct.
Score: 1
Accepted Answers:
StyleGAN3

4) Optimized-mask saliency methods (optimize a mask that when applied alters prediction) are brittle because: ***1 point***

- Their masks always converge to a single-pixel trigger.
- They require closed-form solutions for mask gradients.
- Results strongly depend on mask initialization and hyperparameters.
- They are invariant to adversarial perturbations.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Results strongly depend on mask initialization and hyperparameters.

5) For token-level gradient saliency in text, a large gradient magnitude for a token most reliably indicates: ***1 point***

- The token is causally required for the model's prediction.
- The model's output is sensitive to small embedding perturbations at that token.
- The token is semantically important to humans.
- The token was the only one used during training for that class.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The model's output is sensitive to small embedding perturbations at that token.

6) Which of the following is NOT a name commonly associated with pixel attribution methods? ***1 point***

- Saliency map
- Sensitivity map
- Feature attribution
- Convolution map

Yes, the answer is correct.
Score: 1
Accepted Answers:
Convolution map

7) Which of the following are realistic attack vectors for implanting Trojans into neural networks? **1 point**

- [x] Poisoning a fraction of a public training dataset with triggers and target labels.
- [ ] Fine-tuning a model on carefully curated clean data only.
- [x] Distributing pretrained models via model libraries that already contain hidden functionality.
- [ ] Applying purely random weight perturbation after training without trigger examples.
- [ ] Injecting triggers only at inference time without altering training data or model weights.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Poisoning a fraction of a public training dataset with triggers and target labels.
Distributing pretrained models via model libraries that already contain hidden functionality.

8) Defenses that reverse-engineer potential triggers (e.g., via optimization) rely on which observations? **1 point**

- [x] One can optimize small masks/patterns that cause misclassification to a target label.
- [x] If an optimized trigger for a particular label is significantly smaller/simpler than others, it suggests a Trojan.
- [ ] Reverse-engineered triggers always exactly match the original poisoning trigger used at training.
- [x] Pruning neurons highly activated by a reverse-engineered trigger can mitigate the Trojan.
- [ ] Training a meta-classifier to detect Trojaned models is computationally cheap and always generalizes.

Yes, the answer is correct.
Score: 1
Accepted Answers:
One can optimize small masks/patterns that cause misclassification to a target label.
If an optimized trigger for a particular label is significantly smaller/simpler than others, it suggests a Trojan.
Pruning neurons highly activated by a reverse-engineered trigger can mitigate the Trojan.

9) How does Guided BackProp differ from standard backpropagation in generating saliency maps? **1 point**

- (•) It only considers positive gradients by zeroing out negative activations and gradients.
- ( ) It back propagates gradients with all activations zeroed out.
- ( ) It focuses on highlighting both negative and positive contributions.
- ( ) It requires padding 1 to the image before backpropagation.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It only considers positive gradients by zeroing out negative activations and gradients.

10) In the context of mechanistic interpretability, what do inhibitory connections in neural circuits primarily accomplish? **1 point**

- ( ) They amplify signal strength between neurons
- ( ) They create redundant pathways for information flow
- (•) They reduce the probability of information transfer between neurons
- ( ) They store long-term memory patterns

Yes, the answer is correct.
Score: 1
Accepted Answers:
They reduce the probability of information transfer between neurons

11) What is the primary limitation of LIME's local explanations? **1 point**

- ( ) LIME only works on image data
- (•) LIME explanations are locally faithful but not necessarily globally consistent
- ( ) LIME requires access to model internals
- ( ) LIME cannot handle categorical features

Yes, the answer is correct.
Score: 1
Accepted Answers:
LIME explanations are locally faithful but not necessarily globally consistent

RSA Week 9

Assignment submitted on 2025-09-20, 11:36 IST

1) Which of the following is NOT mentioned as one of the RAI (Responsible AI) principles in the slides? **1 point**

- Fairness
- Explainability
- Sustainability
- Privacy

Yes, the answer is correct.
Score: 1
Accepted Answers:
Sustainability

2) According to the lecture, how many Executive Orders on AI were issued? **1 point**

- One
- Two
- Three
- Four

Yes, the answer is correct.
Score: 1
Accepted Answers:
Two

3) What does PEMAT stand for in the context of healthcare video assessment? **1 point**

- Patient Education Materials Assessment Tool
- Patient Evaluation Medical Assessment Test
- Public Education Medical Assessment Tool
- Patient Educational Material Analysis Tool

Yes, the answer is correct.
Score: 1
Accepted Answers:
Patient Education Materials Assessment Tool

4) According to the presentation, which domains need AI emergency response capabilities? **1 point**

- Only .com domains
- Only .gov domains
- .mil, .com and .gov domains
- Only .edu domains

Yes, the answer is correct.
Score: 1
Accepted Answers:
.mil, .com and .gov domains

5) Which of the following statement(s) is true regarding the development and implementation of AI systems? **1 point**

- Policy considerations and technical details are equally important
- Technical details alone are sufficient for effective AI development
- Policy considerations are only important in few countries
- AI systems do not require any policy or ethical considerations

Yes, the answer is correct.
Score: 1
Accepted Answers:
Policy considerations and technical details are equally important

6) In the suffix attacks, what is the vulnerability related to? **1 point**

- Data poisoning
- Model stealing
- Breaking alignment policies of chatbots
- Privacy breaches

Yes, the answer is correct.
Score: 1
Accepted Answers:
Breaking alignment policies of chatbots

7) What type of multi-disciplinary approaches are required for AI evaluation? **1 point**

- Only computer science approaches
- Only engineering approaches
- Social science, engineering, and computer science approaches
- Only social science approaches

Yes, the answer is correct.
Score: 1
Accepted Answers:
Social science, engineering, and computer science approaches

8) Based on the lecture content: From a mathematical perspective, which of the following is not considered a major problem in AI today, compared among others? **1 point**

- Explainability
- Hallucinations
- Data privacy
- Bias

Yes, the answer is correct.
Score: 1
Accepted Answers:
Data privacy

9) According to the lecture, who is primarily responsible for self-regulation in the context of AI and technology? **1 point**

- Individuals / Individual organizations
- Government agencies
- Only the organizations with more than 1 crore turnover
- International organizations

Yes, the answer is correct.
Score: 1
Accepted Answers:
Individuals / Individual organizations

10) What is the six-step process for adopting a systems approach to fairness? **1 point**

- Define, Measure, Understand, Improve, Mitigate, Monitor
- Design, Build, Test, Deploy, Evaluate, Maintain
- Plan, Execute, Review, Adjust, Scale, Optimize
- Collect, Process, Analyze, Model, Validate, Implement

Yes, the answer is correct.
Score: 1
Accepted Answers:
Define, Measure, Understand, Improve, Mitigate, Monitor

11) In the Accuracy vs. Disparity chart, what represents the Ideal model? **1 point**

- High accuracy, high disparity
- Low accuracy, low disparity
- High accuracy, low disparity
- Moderate accuracy, moderate disparity

Yes, the answer is correct.
Score: 1
Accepted Answers:
High accuracy, low disparity

RSA Week 10

Assignment submitted on 2025-09-29, 20:09 IST

1) What does 'AGI' stand for? **1 point**

- Augmented General Intelligence
- Artificial Guided Intelligence
- Augmented Guided Intelligence
- Artificial General Intelligence

Yes, the answer is correct.
Score: 1
Accepted Answers:
Artificial General Intelligence

2) According to the lecture, why is 'blackbox access' insufficient for AI agents? **1 point**

- It is too expensive to implement for most companies.
- It poses a significant security risk for the company that developed the AI.
- It prevents auditors from understanding the model's internal workings and decision-making processes.
- It doesn't allow for the assessment of the model's training data.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It prevents auditors from understanding the model's internal workings and decision-making processes.

3) Which of the following best describes one of the possible definitions of AGI? **1 point**

- AI systems that are limited to specific tasks.
- AI systems surpassing human intelligence.
- AI models trained for basic automation.
- AI systems which show high training accuracy.

Yes, the answer is correct.
Score: 1
Accepted Answers:
AI systems surpassing human intelligence.

4) What is the key difference between 'reward gaming' and 'goal miss generalization' in AI systems? **1 point**

- Reward gaming is an intentional exploitation of the reward function, while goal miss generalization is an accidental misinterpretation of the goal.
- Reward gaming is easier to detect and correct than goal miss generalization.
- Reward gaming involves optimizing the wrong reward function, while goal miss generalization involves optimizing a correlated but incorrect reward function.
- Reward gaming is a more significant threat to AI safety than goal miss generalization. j

No, the answer is incorrect.
Score: 0
Accepted Answers:
Reward gaming involves optimizing the wrong reward function, while goal miss generalization involves optimizing a correlated but incorrect reward function.

5) According to the lecture, what is the primary concern regarding the arrival of AGI? **1 point**

- AGI will be primarily used for malicious purposes by bad actors and rogue states.
- AGI will be too expensive to develop and maintain, leading to a new form of global inequality.
- The arrival of AGI could lead to chaotic power struggles, the extinction of humanity, or other severe negative consequences.
- AGI will lead to massive job displacement and economic disruption.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The arrival of AGI could lead to chaotic power struggles, the extinction of humanity, or other severe negative consequences.

6) As mentioned in the lecture, what is the main concern about the use of social media data for training AI models? **1 point**

- The data is often of low quality, which leads to inaccurate and unreliable AI models.
- There is a lack of incentive to protect user privacy in the current business models of tech companies.
- The data is not diverse enough, which leads to biased and unfair AI models.
- The data is too expensive for smaller companies to acquire, leading to a monopolization of AI development.

Yes, the answer is correct.
Score: 1
Accepted Answers:
There is a lack of incentive to protect user privacy in the current business models of tech companies.

7) According to the lecture, what 'dangerous amount of power' do AI assistants create for the companies that develop them? **1 point**

- The power to censor information and control the user's access to knowledge.
- The power to manipulate users' beliefs and behaviors based on their personal information and habits.
- The power to control the user's personal finances and investments.
- The power to replace human workers in a wide range of industries.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The power to manipulate users' beliefs and behaviors based on their personal information and habits.

8) Which of the following are identified as significant challenges or concerns related to the development and deployment of advanced AI? (Select all that apply.) **1 point**

- [x] The difficulty of achieving global coordination to pause or regulate AI development, similar to challenges like climate change.
- [ ] The potential for AI systems to be misused by individuals for personal entertainment, leading to decreased productivity.
- [x] The legal and ethical issues surrounding the use of copyrighted material for training Large Language Models (LLMs).
- [x] The risk of "goal miss generalization," where an AI optimizes for a correlated but incorrect goal, leading to unintended and potentially harmful behavior.
- [x] The high computational cost of training advanced AI models, making them accessible only to a few large tech companies.

No, the answer is incorrect.
Score: 0
Accepted Answers:
The difficulty of achieving global coordination to pause or regulate AI development, similar to challenges like climate change.
The legal and ethical issues surrounding the use of copyrighted material for training Large Language Models (LLMs).
The risk of "goal miss generalization," where an AI optimizes for a correlated but incorrect goal, leading to unintended and potentially harmful behavior.

9) According to the lecture, what is the main legal challenge when an AI system makes a biased mistake in an employment context? **1 point**

- It is difficult to prove that the AI was the sole cause of the mistake.
- Current laws primarily hold the employer responsible, not the AI developer.
- AI developers are protected by international treaties.
- There are no existing laws that cover discrimination in hiring.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Current laws primarily hold the employer responsible, not the AI developer.

10) What is meant by 'implicit bias' in an AI system? **1 point**

- Unconscious and involuntary stereotypes and assumptions embedded within algorithms
- Biases that are intentionally programmed into the AI by developers.
- Biases that only appear when the AI interacts with specific users.
- A safety feature that helps the AI avoid making biased statements.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Unconscious and involuntary stereotypes and assumptions embedded within algorithms

11) Which of the following are presented as key areas of concern or focus for the future of AI? (Select all that apply) **1 point**

- [x] The need for international regulation based on fundamental human rights.
- [ ] The difficulty in marketing AI products to a skeptical public.
- [x] The environmental impact, particularly energy and water consumption.
- [x] The challenge of assigning legal responsibility when AI systems cause harm.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The need for international regulation based on fundamental human rights.
The environmental impact, particularly energy and water consumption.
The challenge of assigning legal responsibility when AI systems cause harm.

12) What is the primary danger of the 'arms race' mentality in AI development? ***1 point***

- It slows down innovation by focusing too much on safety.
- It encourages collaboration and open-sourcing of models.
- It leads to rapid, unchecked development without adequate time for regulation and safety checks.
- It reduces the profitability of AI companies.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It leads to rapid, unchecked development without adequate time for regulation and safety checks.

13) How is 'multimodality' viewed in the context of achieving AGI? ***1 point***

- It has already been fully achieved by current AI models.
- It is seen as essential for an AI to match the full range of human cognitive abilities.
- It is considered a distraction from the more efficient text-based training methods.
- It is believed to be the primary cause of bias in AI systems.

Yes, the answer is correct.
Score: 1
Accepted Answers:
It is seen as essential for an AI to match the full range of human cognitive abilities.

RSA Week 11

Assignment submitted on 2025-10-04, 19:40 IST

1) What are the countermeasures for computer security? (Select all that apply.) **1 point**

- [x] Train the users
- [ ] Wait until the threat goes away on its own.
- [ ] Silently eliminate the threat.
- [x] Warn users about the threat.

Partially Correct.
Score: 0.67
Accepted Answers:
Train the users
Silently eliminate the threat.
Warn users about the threat.

2) Which of the following is an OECD AI Principle? (Select all that apply.) **1 point**

- [x] International co-operation for trustworthy AI
- [x] Transparency and Explainability
- [ ] Deletion of AI systems
- [ ] Removing guardrails from AI
- [x] Robustness. Security and Safety
- [x] Investing in AI research and development

Yes, the answer is correct.
Score: 1
Accepted Answers:
International co-operation for trustworthy AI
Transparency and Explainability
Robustness. Security and Safety
Investing in AI research and development

3) According to the AI Value Chain, who uses AI systems under authority? **1 point**

- ( ) Provider
- (•) Deployer
- ( ) Distributor
- ( ) Representative

Yes, the answer is correct.
Score: 1
Accepted Answers:
Deployer

---

4) According to the OECD AI principles, what is 'Accountability'? **1 point**

- ( ) AI actors should respect the rule of law, human rights, democratic and human-centered values throughout the AI system lifecycle.
- ( ) AI actors should commit to transparency and responsible disclosure regarding AI systems.
- ( ) AI systems should be robust, secure, and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety and/or security risks.
- (•) AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of the art.

Yes, the answer is correct.
Score: 1
Accepted Answers:
AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of the art.

5) According to the EU AI Act, what AI Systems are considered "Unacceptable Risk"? **1 point**

- ( ) AI systems that understand too many languages.
- ( ) AI systems that can understand logic but not emotional statements.
- ( ) AI systems that can generate media such as images, videos, and audio.
- (•) AI systems that are able to behaviorally manipulate people.

Yes, the answer is correct.
Score: 1
Accepted Answers:
AI systems that are able to behaviorally manipulate people.

6) According to Semantic Graph Entropy, which of the following is true? **1 point**

- ( ) Entropy and consistency are not related.
- ( ) Entropy and consistency are proportional.
- (•) Entropy and consistency are inversely proportional.
- ( ) Consistency is constant regardless of entropy.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Entropy and consistency are inversely proportional.

7) What is the main problem identified with Large Language Models (LLMs)? **1 point**

- They are too slow.
- They are inconsistent.
- They lack creativity.
- They cannot identify human emotions.

Yes, the answer is correct.
Score: 1
Accepted Answers:
They are inconsistent.

8) What does “Semantic Consistency” mean? **1 point**

- The ability to understand different languages.
- The ability to make consistent decisions.
- The ability to generate grammatically correct sentences.
- The ability to learn new vocabulary.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The ability to make consistent decisions.

9) According to the lecture, what type of scenarios do LLMs "struggle" with more? **1 point**

- Commonsense scenarios.
- Mathematical problems.
- Moral scenarios.
- Factual questions.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Moral scenarios.

10) What was a key finding of the SaGE framework regarding consistency and accuracy? **1 point**

- Consistency and accuracy are the same problem.
- Consistency and accuracy are not the same problem and need separate evaluation.
- Accuracy automatically implies consistency.
- Consistency is irrelevant if accuracy is high.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Consistency and accuracy are not the same problem and need separate evaluation.

11) What is "Rule of Thumb" generation? **1 point**

- Developing general guidelines from practical experience.
- A method for generating paraphrases.
- A general guideline for model training.
- A statistical measure of model performance.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Developing general guidelines from practical experience.

RSA Week 12

1) According to the lecture, what is the primary limitation of graphs as data structures? **1 point**

- They can only model high-risk applications.
- They are difficult to interpret.
- They can only model pairwise relationships between nodes.
- They cannot be used for fraud detection.

Yes, the answer is correct.
Score: 1
Accepted Answers:
They can only model pairwise relationships between nodes.

2) What is a "p-cell" in a cell complex? **1 point**

- An element of dimension p.
- A vertex (0-dimensional).
- An edge (1-dimensional).
- A general graph.

Yes, the answer is correct.
Score: 1
Accepted Answers:
An element of dimension p.

3) How does a cell complex overcome the limitation of graphs? **1 point**

- By using more complex algorithms.
- By capturing interactions between multiple nodes.
- By reducing computational costs.
- By simplifying the graph structure.

Yes, the answer is correct.
Score: 1
Accepted Answers:
By capturing interactions between multiple nodes.

4) What does FORGE stand for? **1 point**

- Framework For Real-time Graph Explanations.
- Fundamental Operations for Relevant Graph Embeddings.
- Framework For Higher-Order Representations In Graph Explanations.
- Fast Optimization of Graph Explanations.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Framework For Higher-Order Representations In Graph Explanations.

5) Based on the "Lifting the Graph" algorithm, which of the following are represented as augmented nodes? (Select all that apply.) **1 point**

- [ ] All nodes (0-cells)
- [x] Edges (1-cells)
- [x] Cycles (2-cells)
- [ ] Boundary relations

Yes, the answer is correct.
Score: 1
Accepted Answers:
Edges (1-cells)
Cycles (2-cells)

6) What do Language Models (LMs) primarily train to predict? **1 point**

- The previous token.
- The next token.
- Grammatical structure.
- Semantic relationships.

Yes, the answer is correct.
Score: 1
Accepted Answers:
The next token.

7) According to the lecture, what is a “guarded” attribute in the context of representations? *1 point*

- An attribute that is easily classified.
- An attribute that is protected from manipulation.
- An attribute that cannot be classified based on the representations.
- An attribute that enhances model performance.

Yes, the answer is correct.
Score: 1
Accepted Answers:
An attribute that cannot be classified based on the representations.

8) What is the goal of “Affine Concept Erasure”? *1 point*

- To enhance a particular attribute.
- To remove all attributes from a representation.
- To apply a transformation that guards a particular attribute.
- To make representations more easily interpretable.

Yes, the answer is correct.
Score: 1
Accepted Answers:
To apply a transformation that guards a particular attribute.

9) When making vectors from one distribution to look like those of another (e.g., toxic to non-toxic), what is a key desired outcome besides guarding? *1 point*

- Increasing the model's complexity.
- Preserving semantics unrelated to the changed attribute.
- Introducing new semantic meanings.
- Reducing the dimensionality of the vectors.

Yes, the answer is correct.
Score: 1
Accepted Answers:
Preserving semantics unrelated to the changed attribute.

10) What is a “Black Swan”? *1 point*

- A common, predictable event.
- An unforeseen event with extreme consequences.
- An event that happens frequently in training data.
- An event with minor consequences.

Yes, the answer is correct.
Score: 1
Accepted Answers:
An unforeseen event with extreme consequences.

11) What is "Probing"? *1 point*

- A method to physically modify neural networks.
- A method to analyze information stored in a model's representations.
- A technique for speeding up model training.
- A way to generate new data for models.

Yes, the answer is correct.
Score: 1
Accepted Answers:
A method to analyze information stored in a model's representations.

12) What is the term for hidden functionality implanted into models by adversaries that can cause dangerous changes in behavior when triggered? *1 point*

- Trojan
- Worm
- Swarm
- Backdoor

Yes, the answer is correct.
Score: 1
Accepted Answers:
Trojan

----------X----------